frontiers | Frontiers in Computer Science

TYPE Original Research
PUBLISHED 29 September 2025
DOI 10.3389/fcomp.2025.1575296

OPEN ACCESS

EDITED BY
Stefania Serafin,
Aalborg University Copenhagen, Denmark

REVIEWED BY
Constance Bainbridge,
University of California, Los Angeles, United States
Gerardo Acosta Martínez,
University of York, United Kingdom
M Fahim Ferdous Khan,
Toyo University, Japan

*CORRESPONDENCE
Camille Noufi
✉ cnoufi@ccrma.stanford.edu

RECEIVED 12 February 2025
ACCEPTED 05 September 2025
PUBLISHED 29 September 2025

CITATION
Noufi C, May L and Berger J (2025) A model of vocal persona: context, perception, production. *Front. Comput. Sci.* 7:1575296. doi: 10.3389/fcomp.2025.1575296

A model of vocal persona: context, perception, production

Camille Noufi*, Lloyd May and Jonathan Berger

Center for Computer Research in Music and Acoustics (CCRMA), Stanford University, Stanford, CA, United States

We present a contextualized production–perception model of vocal persona developed through deductive thematic analysis of interviews with voice and performance experts. Our findings reveal that vocal persona is a dynamic, context-responsive set of vocal behaviors that frames and bounds expressive interactions—both biological and synthesized—while centering the speaker's agency. By examining how experts adapt their vocal output through both broad persona shifts and fine-grained paralinguistic adjustments, our model identifies a key missing mechanism in current approaches to expressive speech synthesis: the integration of high-level persona prompting with detailed paralinguistic control. This work bridges an important gap in the literature on expressive and interactive speech technologies and offers practical insights for improving voice user interfaces and augmentative and alternative communication systems. Incorporating this vocal persona framework into expressive speech synthesis holds the potential to enhance user agency and embodiment during communication, fostering a heightened sense of authenticity and a more intuitive relationship with voice interaction technology and one's environment.

KEYWORDS

vocal persona, social communication, expression, paralinguistics, synthesized voice, augmentative and alternative communication (AAC), voice user interface (VUI)

1 Introduction

The concept of vocal persona, put forth by musicologist Phillip Tagg, encapsulates the intricate relationship between the voice and our identities, emotions, and behavioral positions (Tagg, 2012). Tagg's description, rooted in the Latin term "persona" (literally "heard through"), moves beyond its theatrical origins to encompass the everyday roles and expressions adopted by individuals. This concept emphasizes that personas, which dynamically adapt and respond to the surrounding environment and context, are not limited to performers but extend to all individuals in various life situations (Goffman, 2016; Marshall and Barbour, 2015). It distinctively narrows the focus to the domain of vocal expression and emphasizes the voice as a key medium for persona manifestation.

The concept of a vocal persona is well recognized among performers, voice coaches, and voice technology experts (Linklater, 2006), yet it remains largely underdefined in academic research. Current industry practices often prompt a persona using high-level character descriptions, allowing the synthesizer to generate an appropriate voice (Guo et al., 2022; Shimizu et al., 2024; Yang et al., 2024). However, these approaches lack a mechanism for fine-grained paralinguistic control—namely, the ability to adjust non-verbal vocal cues such as tone, pitch, and rhythm (Schuller et al., 2013)—that is essential for shaping a dynamic, contextually appropriate vocal persona. In other words, while a persona can be prompted, there is no systematic method to introduce precise variability or to instruct specific paralinguistic modifications once that persona is established. This gap

is especially pertinent in synthetic speech technologies for Augmentative and Alternative Communication (AAC) systems, where incorporating refined, controllable vocal attributes could lead to more expressive and personalized outputs. Bridging the expertise from the performing arts with advances in Human-Computer Interaction (HCI) and Voice User Interface (VUI) research holds significant promise, particularly for these systems. Such systems have long been critiqued for their lack of contextualized expressivity and personalization (Pullin and Hennig, 2015), underscoring the need for a unified framework that enables both high-level persona prompting and precise paralinguistic modulation.

Our study addresses this gap by examining how vocal professionals utilize precise paralinguistic control to shape their vocal personas in human-to-human communication. We sample voice professionals as power users whose tacit expertise makes explicit the levers of vocal persona, providing a design vocabulary that can be selectively adapted for less expert populations. Rather than focusing on algorithmic or programmatic solutions, our work prioritizes understanding the embodied practices that underlie vocal persona formation and modulation. Specifically, our research aims to answer the following questions:

1. What are the defining characteristics of a vocal persona as articulated by vocal professionals, and how do they implement paralinguistic modification to refine or adjust this persona?
2. Is there a structured relationship between context, vocal expression, and identity that governs vocal persona in both natural interactions and VUI?
3. In what ways can an enhanced understanding of human vocal persona dynamics inform the design of more expressive and personalized VUI systems?

To explore these questions, we conducted in-depth interviews with twenty-one vocal professionals to gain insights into the interplay between vocal expression, identity, environmental cues, and user feedback. The rich insights derived from these discussions reveal how vocal personas are dynamically crafted and adjusted in everyday communication. Building on these findings, we developed a model of vocal persona that formalizes the interplay between identity, context, and vocal expression—with a particular focus on the role of paralinguistic modifications. This model provides a structured framework for understanding vocal persona dynamics in human interactions and lays the groundwork for future research aimed at translating these insights into improved expressive speech synthesis and VUI systems. Our approach is intentionally qualitative, prioritizing depth and the surfacing of latent structure over quantitative generalizability. The resulting model is intended as a design vocabulary and conceptual framework, not as a validated end-user tool. Quantitative and behavioral validation of the model's applicability to broader user populations is an important direction for future work.

2 Background

Effective communication depends not only on the words we choose but also on how our voices are modulated by context. Traditional models such as Berlo's unidirectional SMCR (Sender-Message-Channel-Receiver) (Berlo, 1960) have proven inadequate by treating communicators as passive channels. In contrast, dynamic frameworks like the Interactive and Transaction models emphasize that communication is a co-creative process, continuously shaped by physical, psychological, and cultural influences (Westley and Malcolm S. MacLean, 1957; Barnlund, 1970). This reconceptualization establishes a foundation for understanding how nuanced vocal modulation serves as a critical tool for both personal expression and social interaction.

Building on this foundation, research into vocal identity reveals that the voice is not merely a medium for conveying words but a performative tool for negotiating and presenting one's identity. Goffman's insights into self-presentation (Goffman, 2016) and Butler's theory of gender performativity (Butler, 1990) illustrate that vocal behavior is dynamic and context-dependent. Speakers routinely engage in code-switching—altering intonation, rhythm, and accent—to signal shifts in social roles and identity (Bullock and Toribio, 2009). Moreover, recent findings by Guldner et al. (2024) show that speakers deliberately modulate their vocal expressions to accentuate traits such as confidence and likeability, aligning these adjustments with perceptual dimensions of affiliation and competence. Together, these perspectives make it clear that precise, context-driven paralinguistic modulation is central to constructing an authentic vocal persona, thereby bridging internal identity and public performance.

Beyond identity, the subtleties of vocal expressivity encompass the transmission of emotion and personality. Acoustic features such as timbre, fundamental frequency, and intonation play a pivotal role in conveying emotional arousal and pleasantness (Bachorowski, 1999). Early studies established correlations between vocal traits and perceived personality—highlighting influences like gender (Addington, 1968) and linking extroversion to dynamic vocal effort (Scherer, 1978). Later research has mapped personality impressions onto dimensions of Valence and Dominance (McAleer et al., 2014), while studies by Stern et al. (2021) and Nass et al. (1994, 1995, 2001) reveal that even subtle paralinguistic cues trigger robust, often unconscious, social responses. These findings underscore that nuanced vocal modulation is indispensable not only for conveying identity but also for authentically transmitting emotion and personality.

In today's landscape—where generative models like GPT-4 (OpenAI et al., 2024) and advanced neural Text-to-Speech (TTS) systems are transforming expressive agent capabilities—significant progress in speech synthesis has been made. Yet, these current systems still exhibit a disjoint between high-level persona prompting and fine-grained paralinguistic control. On one side, techniques such as those proposed in PromptTTS (Guo et al., 2022) demonstrate high-level persona prompting—using character descriptions or other widely-understood contextual cues—to evoke a desired vocal identity. On the other side, fine-grained control has been realized through example-based approaches (Wang et al., 2018; Valle et al., 2020b; Shechtman et al., 2021; Wang et al., 2023) and direct manipulation methods (Sorin et al., 2017; Wang et al., 2018; Hsu et al., 2019; Valle et al., 2020a; Morrison et al., 2021b,a; Neekhara et al., 2021) that adjust specific acoustic features. However, these two streams still remain largely separate, and no

unified framework has yet been developed to integrate high-level persona prompting with precise paralinguistic adjustments—a gap that highlights the contrast between the synthesized vocal personas in industry with the nuanced control seen in natural human communication.

To address this gap, our study focuses on understanding how vocal professionals embody and manipulate their vocal personas across a broad spectrum of social interactions and technology-mediated environments. By examining the interplay between contextual cues, identity, and fine-grained paralinguistic modulation, we aim to develop a framework that unifies high-level persona prompting with situational vocal control. This framework aims to improve our understanding of natural vocal behavior in a wide variety of contexts and can inform future enhancements in expressive speech synthesis, Voice User Interfaces, and Augmentative and Alternative Communication systems.

3 Methods

3.1 Study design and setup

We conducted a deductive thematic analysis to examine the qualitative data collected from semi-structured interviews with voice professionals, beginning with an a priori framework and iteratively refining codes through multiple rounds of independent coding and consensus reconciliation. This methodology allowed for the incorporation of predefined and emergent themes, facilitating a contextual understanding of the interview data (Terry et al., 2017). To enhance trustworthiness, we employed iterative coding with collaborative reconciliation; coders discussed discrepancies in code application and adjusted the codebook until full consensus was reached (Nowell et al., 2017).

The study received Institutional Review Board (IRB) approval from Stanford University prior to data collection. Participants were recruited through email listservs related to voice, language, performing arts, and communications. Semi-structured interviews were conducted over the Zoom video conferencing platform using a set of predetermined a priori interview questions. Participants were recruited and compensated for their time.

Initial themes and questions were created through an iterative process of literature review and discussion among the research team. The generated questions were then clustered, forming the initial themes and sub-themes, namely: Physical Context, Environment and Space, Sociocultural Context, The Role of Technology, Self-Perception and Perception of Others, and Agency. A complete list of themes, sub-themes, and interview questions is available in the Supplementary Materials.

3.2 Participants

The study involved twenty-one adults with experience and expertise in a voice-related profession who participated in semi-structured interviews. Before the interviews, the participants were asked to complete a pre-interview survey that gathered information about their primary and secondary profession expertise, age, gender identity, multilingualism, and English language comprehension, summarized in Table 1.

The participants' ages ranged from 22 to 67 years (median = 39, IQR = 21), and 62% reported being multilingual. Nine (42.9%) participants identified as she/her, ten (47.6%) identified as he/him, and two (9.5%) did not report their gender identity. The primary experiences reported were: Singing/Vocal Performance (10), Voice Technology & Science (4), AAC User (3), Acting/Voice-Over Work (2), Linguistics (2), and Poetry/Songwriting (1). Pedagogy/Teaching (7) was the most common secondary experience type reported, followed by Singing/Vocal Performance (2), Voice Technology & Science (2), Linguistics (1), Acting/Voice-Over Work (1), and Journalism (1). Five participants did not report secondary experience. The participants had between 2 to 51 years of primary experience (median = 15, IQR = 12) and 2 to 36 years of secondary experience (median = 16.5, IQR = 19.25). We targeted approximately 50% of our participants to be performers as their lived experience and subsequent relationship to persona is under-explored.

3.3 Data collection and analysis

The data collection process involved twenty-one interviews conducted by two researchers, with each researcher interviewing approximately half of the participants. The goal of these interviews was to surface how trained vocal professionals conceive, construct, and deploy vocal personas; their expertise shapes their awareness and performance, making the sample a purposeful, bounded source of deep insight rather than a representative population. The interviews followed a semi-structured format, where participants were encouraged to discuss topics that were most relevant to them. This approach allowed participants to express their thoughts and experiences freely, which aided the researchers in gaining insights into a priori and emergent themes.

The interviews were transcribed using Zoom's automatic transcription, manually cleaned, then anonymized. The transcripts, video, and audio of the interviews were then uploaded to secure storage. Coding and analysis was performed using Dedoose Version 9.0.82.

In the first round of coding, each researcher coded interviews they did not participate in, using both a priori themes and emerging themes. When categorizing excerpts, the functional emphasis in the participant's account was used as the primary decision criterion for placement within the current round's thematic structure. Changes were documented in the Change Log.[1] This was done in two parts: initial coding of 10 interviews using a priori themes and sub-themes, followed by intermediate analysis with emerging codes and themes. Another 11 interviews were then coded. In the second round, the original interviewer reviewed the coded excerpts, considering modifications from the first round, refining and consolidating the coding scheme for consistency, and making adjustments to the codebook. A third round of collaborative analysis was conducted to ensure agreement on

1 A priori and emerging themes, as well as the Change Log, can be found in the Supplementary materials.

TABLE 1 Participant self-reported demographic information and experience related to voice and/or communication.

Identifier	Age (years)	Gender identity	Primary experience	Years (primary)	Secondary experience	Years (secondary)	Multilingual
1	25	She/her	Linguistics	2	Singing/vocal performance	*NR*	Yes
2	59	He/him	Acting/voice-over work	51	Pedagogy/teaching	26	Yes
3	50	He/him	Singing/vocal performance	30	Acting/voice-over work	30	Yes
4	67	She/her	Singing/vocal performance	30	Pedagogy/teaching	26	No
5	46	She/her	Singing/vocal performance	20	Pedagogy/teaching	10	No
6	26	She/her	Singing/vocal performance	14	Voice technology & science	5	Yes
7	33	*NR*	Singing/vocal performance	10	*NR*	*NR*	Yes
8	48	She/her	Voice technology & science	18	Linguistics	18	No
9	39	He/him	Singing/vocal performance	22	*NR*	*NR*	No
10	27	He/him	Poetry/songwriting	15	Singing/vocal performance	15	No
11	68	she/her	Singing/vocal performance	40	Pedagogy/teaching	36	No
12	45	She/her	Singing/vocal performance	30	Pedagogy/teaching	20	No
13	25	She/her	AAC user	15	*NR*	*NR*	Yes
14	30	He/him	Singing/vocal performance	20	Voice technology & science	2	Yes
15	33	He/him	Voice technology & science	2	Singing/vocal performance	7	No
16	22	He/him	AAC user	10	Journalism	*NR*	Yes
17	52	He/him	Acting/voice-over work	15	Pedagogy/teaching	15	No
18	42	He/him	AAC user	10	*NR*	*NR*	Yes
19	40	*NR*	Voice technology & science	5	*NR*	*NR*	Yes
20	30	She/her	Singing/vocal performance	10	Pedagogy/teaching	6	Yes
21	25	He/him	Voice technology & science	3	*NR*	*NR*	Yes

NR indicates information not reported on the pre-interview demographic and experience survey.

all coded excerpts and modify codes/themes as needed, before finalizing the codebook.

4 Results

Three overarching themes emerged from the analysis: Context, Production, and Perception. The codebook found in the Supplementary materials provides definitions of the themes, subthemes, and codes, generated from the thematic analysis. Below we describe each of the three main themes and their relationship to vocal persona.

While certain examples could plausibly fit more than one theme, we assigned them based on the function most foregrounded in the participant's account. Items described as external conditions or constraints shaping vocal choices were placed under Context, those described as active, intentional modulation strategies under Production, and those focused on interpretation or feedback loops under Perception. This functional emphasis guided placement in cases where phenomena might span multiple categories. For example, masks, makeup, or costumes may be seen as contextual or perceptual, but were coded under Production when participants emphasized their role in actively altering embodiment and vocal delivery.

4.1 Context

Four forms of context were found to influence the choice of a vocal persona: physical, sociocultural, temporal, and technological. In the study, participants revealed that these factors all significantly impact their choices during communication. They adjusted their vocal production for specific locations, such as using varied tones in restaurants versus their living rooms. Ambient noise, noise changes, and the social acceptance of noise (for example, a church vs. a lecture hall versus a concert hall) influenced their persona selection. To ensure intelligibility in noisy or reverberant environments, participants either chose assertive or authoritative personas or accentuated key vocal characteristics of their current persona. They reported using between 2 to 7 different vocal personas tailored to various specific contexts (median $=$ 5, IQR $=$ 4.5). The time of day and week also significantly affected their vocal choices. Participants indicated that long-term contextual variations often determined the selection from a broader range of personas, whereas short-term variations typically influenced how they utilized vocal characteristics within a chosen persona.

Sociocultural context also emerged as a powerful influence. Many participants referred to their voice as an outward-facing communication channel of a personality. Participants mentioned

adjusting their vocal persona to orient toward a social relationship between 2 and 5 times per day (median = 3, IQR = 2.5). P15 stated, *"Having different voices for different scenarios is very much akin to how we have different personalities. It's like a totally different base for interacting..."*. The vocal embodiment of a persona often involves acquired paralinguistic mimicry of an archetype or character, with the successful interpretation of this vocal orientation requiring agreement among all parties in the conversation. This agreement involved (1) understanding the vocal characteristics of the character, (2) agreeing that the vocal cues portrayed by the speaker were correct in orientation, and (3) believing that the degree to which this orientation occurred was desirable and correct. For example, P20, a professional singer and university teaching assistant, consciously adopts an "authoritative educator" vocal persona during her teaching sections. She utilizes specific acoustic paralinguistic cues: her voice is modulated to have a clear tone, a steady pace, and intentional pauses to convey emphasis, confidence, and knowledge. This vocal portrayal is recognized by her students, who understand her clear enunciation and assertive tone as embodying an educator's role. There's a mutual agreement that these vocal characteristics are apt for teaching, with the firm yet non-threatening tone fitting the educational context. Both P20 and her students find this vocal orientation desirable, as it fosters a respectful and focused learning environment: *"[I] get a lot of feedback from my students that they really feel like they know that I know what I'm talking about, and I think a lot of that comes from the confidence... in my voice."* This persona adoption differs from P20's approach to teaching private voice lessons: *"I try to speak as if you're speaking to a friend. I try to adopt a really approachable persona.... but I think a lot of [those cues] are unconsciously done."*

The adoption of a vocal persona was also influenced by one's orientation toward a specific language or culture. Multilingual participants described how sociocultural norms shaped not just vocabulary or accent, but modulation of specific paralinguistic characteristics of their voice. This finding is discussed further in Section 4.2.

Technology was the fourth major contextual influence, and participants described it as uniquely capable of altering both the conditions of communication and the range of available expressive strategies. The influence of technology on vocal expression is a relatively new phenomenon as it affects the way we vocally orient and respond to context. Technology was found to be a unique context that allowed participants to (1) alter the acoustics within the transmission and/or feedback of information and (2) separate vocal embodiment from other modalities of embodiment. Technology influenced not just the transmission of the vocal signal but also affected vocal style. The use of microphones shifted singers' technique choices toward expression and style over intelligibility. For example, microphones and amplification allowed for unique types of speech, such as loud, intelligible whispers, that emulated physical contexts like close proximity. P4 explained, *"This allows for more intimacy and softness in singing. Without a microphone, singing intimately still requires maintaining a strong center [to create resonance]. However, when singing with a microphone, you can occasionally relax this requirement... You can get away with a broader range [of vocal expressions], including that style of breathy tones and breathy pickups..."*

Vocal effects and other technological manipulations influenced both the adoption and self-perception of a vocal persona. For example, singers discussed how the use of different styles of microphones, such as lapel mics attached to the performer's clothes versus a hand-held mic, influenced their vocal persona. Participants also discussed how communication occurring exclusively through an audio channel often resulted in more nuanced or dramatic use of paralinguistic cues, influencing the variability and range of the paralinguistic attributes highlighted within a vocal persona, as compared to within an audiovisual setting. Participants also acknowledged how technological advancements over time have affected the choice of persona employed during interpersonal interactions, such as the increased occurrence of phone conversations over face-to-face conversations. P15 noted how his voice has become quieter during arguments due to the mediation of phone and video technology, and he has noticed this transition over several years affecting the vocal persona he adopts when arguing in person.

4.2 Production

The second theme focuses on how vocal persona influences vocal production. It demonstrates that adopting a specific persona directly affects the manipulation and prioritization of paralinguistic attributes in voice production. This theme encompassed four main sub-themes: vocal characteristics, intention and planning, embodiment, and self-initiated synthesis.

Participants described the complex relationship between their internal state and outward expression, highlighting how intentional manipulations of paralinguistic attributes could be used to control how emotions or intentions were perceived. The adoption of a vocal persona allowed them to act upon their communication priorities in context, providing instances of both intentional and unintentional disconnecting between internal state and outward expression. For instance, some reported feeling angry but having to communicate calmly in a professional setting. Performers spoke of being coached to embody high-arousal emotions intentionally or to not feel the emotion at all but, instead, to intentionally manipulate particular paralinguistic attributes to project it to an audience. P3 describes this intention when singing as a character in an opera: *"I'm not trying to feel pain, or get you to feel pain, I need to communicate to you that [my character is] in pain."*" Instances of disconnect were influenced by sociocultural and physical context, and, when done intentionally, were achieved through self-aware, deliberate modification of characteristics of the voice. Performers seemed to be particularly attuned to how a vocal persona would be perceived and described exactly what modifications they make to their voice and body to achieve their desired effect. Non-performers similarly seemed to have this capability but were less able to describe the process in detail. However, both performers and non-performers noted that the degree of conscious intention and planning was primarily context-dependent. In high-arousal states, such as emergencies, participants reported a more direct connection between internal state and outward expression. These states were vocalized more rapidly, and at times, the act of

expressing paralinguistic elements took precedence over the actual linguistic content.

These production strategies were also shaped by sociocultural and linguistic influences, particularly for multilingual participants. As noted in Section 4.1, these individuals described shifting pitch range, tone, and timbre in ways not dictated by the language itself, resulting in noticeable persona changes. For example, P7, a multilingual participant, noted that his tone and pitch range shift noticeably when speaking different languages: *"I don't talk the same way in English as I do in French. It's about the tone, not just accent... It's something that I feel if I think about it, but it's not something that I [do] consciously."* Although these adjustments were not always deliberate, they reflect the participant's embodied adaptation of paralinguistic attributes in response to sociocultural context. Here, pitch and tone modification are the defining features implemented to shift the speaker's persona. This example also illustrates the interplay between context as a trigger and production as the execution mechanism.

Embodiment emerged as another recurring influence on production choices. Performers noted that masks, makeup, and costumes enhance the distinction between personal identity and outward-facing persona by dissociating their voices from the audience's view of the body's physical appearance. Moreover, performers discussed modification of their physical attributes to better align with the persona of a particular emotion or character.

The analysis also underscored the *prioritization* of intelligibility in choosing a vocal persona and in determining specific paralinguistic attributes. This focus on intelligibility is shaped by both environmental factors, like acoustic feedback, and sociocultural contexts, such as the cultural significance of a performance venue. It's considered vital in various roles, including information delivery, performance settings, and interactions with voice technologies. Importantly, prioritizing intelligibility doesn't detract from stylistic expression; it's achieved by selecting an appropriate vocal persona and, if needed, adjusting paralinguistic attributes. For example, a flight attendant during safety announcements maintains clarity in speech without compromising their friendly and approachable persona.

Prioritization of intelligibility was discussed in particular depth by both performers and AAC device users. The type and degree of mediation of the vocal signal, such as that applied by a phone or vocal microphone, also influence the priority of intelligibility and expression in vocal communication. Participants described feeling the need to modify certain vocal attributes to ensure their intent was conveyed effectively when using a technological mediator, although they made various changes to adhere to other priorities and contexts. Technological interaction with non-human agents, such as screen readers and voice assistants, often prioritized predictability and usefulness over emotional connection, furthering the disconnect between language and paralinguistic cues. P13 noted the effect of this prioritization on his self-perceived social communication abilities: *"I grew up without much mentoring when it came to social skills, and because I listened to a screen reader that is not supposed to have emotion, it's a little harder for me. I also tend to have a habit of speaking in more of a monotone when I'm trying to not display emotion. I have difficulty, for example, being vulnerable around people."*

Finally, several participants discussed the challenges of expressing one's full self when one's voice is impaired or communication abilities are limited. P17 spoke of his inability to vocally express his full range of emotions while ill when consoling his young daughter one evening: *"I wish I could preserve my voice and express emotion and tell [my daughter] that 'it's okay,' you know. To not be able to express your love and even to do simple things like flirt. [Those were] very painful years... in those nine years when I could not communicate, the welling of emotion, it kills you. It really kills you. You cannot imagine how painful it is not to be able to express yourself."*

4.3 Perception

The third theme, perception, encompasses four sub-themes: the perception of others, the perception of voice technologies, multimodal perception, and self-perception.

Participants described how assumptions about a person's abilities and experiences were often made based on their voice, with stereotypes related to gender, race, and nationality shaping these perceptions. They further noted that deliberate and skilled vocal choices could create a more favorable impression, leading to a more generous reading of the speaker's authenticity or credibility. Both short-term and long-term social feedback led to adjustments in participants' vocal personas, with new information or cues prompting changes in vocal behavior. P18, an AAC user, adeptly adjusts his vocabulary to match different social contexts, noting, "You can't use the same verbiage when you are talking with your friends as you do when you are doing a professional presentation. I have different modes of communication for the different positions I occupy." When perceiving others' personas, he placed significant emphasis on paralinguistic cues like "tone, tenure, and volume" and used descriptors such as "stern manner," "elevated tone," "rough," and "smooth and easy going" to discern mood or attitude during vocal communication.

Perceptions also extended to interactions with voice technologies. Participants highlighted that personality traits are often projected onto voice agents, both consciously and subconsciously. They consciously acknowledged designed traits—such as the cheerful tone of a synthesized customer service agent—aimed at improving user experience. Subconsciously, interactions with devices like home voice assistants often led to perceptions of authority, influenced by consistent accuracy and a confident vocal style. This interaction fostered trust and, in some cases, encouraged users to mirror the tone in their own responses. However, this trust proved fragile: erroneous information from the assistant, even in the same confident persona, could rapidly diminish it. In such cases, users often shifted their vocal persona, adopting a more aggressive and agitated tone—a notable departure from their usual human-to-human communication style.

Alongside these perceptions of others and of voice agents, participants reflected on their own voices. While levels of self-awareness varied, all participants described a concept of "my voice"—a personalized range or "sweet spot" within a probability space of possible vocalizations. P4 emphasized, *"If you're playing a character and you're taking on some sort of vocal characteristic,*

you have to be sure that it's still within your own voice, rather than putting on somebody else." The individualized voice was often defined in terms of pitch, cadence, and tone, and was seen as aligned with personal expressive characteristics even when diverging from societal expectations. This perspective revealed a distinction between a "neutral" vocal persona and a neutral vocal tone: the former reflects personality-linked vocal traits, while the latter relates more to physiology. Participants' mental models of how their voice sounded or should sound were often heavily influenced by others' opinions and prevailing norms.

Discussions of self-perception frequently led to distinctions between "normal" and "authentic" vocal personas. While both were often described as "not forced," participants noted that "normal" voices tended to be shaped by societal norms and expectations. In contrast, authenticity was linked to perceived honesty and genuineness in expressing one's thoughts and emotions. The study revealed that naturalness and authenticity could diverge: naturalness referred to the realism of a voice, judged by the absence of unnatural artifacts or violations of vocal expectations, whereas authenticity related to trust in the speaker's intentions. For example, P6 observed that a "natural-sounding" synthesized voice might still convey information in an inauthentic manner, showing that a synthetic quality does not necessarily equate to trustworthiness.

Finally, participants addressed the somatic experience of voice in the body as a central element of self-perception. Many described a disconnect between their physical body and their perception of their own vocalizations when hearing a playback, often due to unexpected timbre or the absence of anticipated physical feedback. This disconnect could result in disembodiment or dissociation from their intended identity. Disembodiment was described as a mismatch between internal state and the information encoded in the vocal signal, arising when the perceived persona of the externally produced voice did not align with the intended persona. P14 noted that karaoke singers sometimes adjusted their performance toward the modified voice if they experienced such disembodiment within a feedback loop: *"... if there's a large enough difference in those two, I think the user actually tries to embody the feedback as something that they can relate to. They decide to use that [modified vocal] feedback to characterize themselves in a way."* They added that this adjustment occurred more often in response to changes in pitch or cadence than to timbre.

Taken together, the themes of context, production, and perception illustrate how vocal persona is both shaped by and shapes communicative interaction. Context provides the triggers and constraints, production executes the physical and expressive strategies, and perception, both internal and external, mediates feedback that informs future adaptation. These interdependencies point toward a dynamic, cyclical model in which all three themes operate in concert rather than in isolation.

4.4 A three-spoke model of vocal persona

Building on the interplay between context, production, and perception, we propose a three-spoke model of vocal persona. This model captures how these dimensions interact in real time to shape both the formation and interpretation of a persona (Figure 1).

The model's first spoke, *Context*, encompasses environmental, cultural, temporal, and technological influences that shape vocal interactions. For instance, a voice assistant might need to modulate its expressiveness when transitioning from a quiet home environment to a noisy public space. The second spoke, *Production*, addresses the technical aspects of creating vocal output, including synthesis technologies, the selection of acoustic features, and the integration of linguistic content. While these two spokes may appear to overlap, they serve distinct roles in persona formation: Context refers to the external circumstances that shape vocal choices—such as audience, setting, cultural norms, and situational goals—whereas Production refers to the speaker's internal modulation strategies, including technical control over vocal parameters like prosody, resonance, and pitch. For example, a performer might adjust pitch height and vowel shaping (Production) when shifting from a classroom to a theatrical setting (Context). Recognizing these as separate but interrelated processes allows the model to accommodate both top-down situational influences and bottom-up vocal strategies.

The third spoke, *Perception*, captures how vocalizations are interpreted by listeners, as well as how they are self-perceived by speakers or conversational agents. It plays a critical role in the iterative refinement of persona. While Context shapes the choices behind vocal construction, Perception reflects how those choices are received, interpreted, and evaluated—and how that feedback may subsequently alter a speaker's strategy. This spoke is bidirectionally linked to persona because perception not only determines how a vocal identity is understood but also shapes how speakers iteratively adjust and embody that identity over time.

4.5 A framework for persona-guided vocalization

Building on our three-spoke model, we propose a modular framework that operationalizes the interplay of context, identity, and paralinguistic control into actionable steps (Figure 2). The process begins with the identification of three preconditions: the speaker's **Intent and Planning**, the prevailing **Context**, and the **Perceptual Feedback** (which includes both self-assessment and environmental cues). For example, during a professional speech, a speaker might aim to project clarity and authority in a formal setting. In this case, the context dictates the selection of a persona that inherently prioritizes such attributes. Within that persona, specific paralinguistic parameters—such as pitch, tone, and rhythm—are then fine-tuned to meet the situational requirements. This cyclical process allows for continuous adaptation: as context or feedback changes, so too can the persona or its paralinguistic details, ensuring that vocal expression remains aligned with communicative goals.

5 Discussion

Our study's key finding was that vocal professionals characterize persona through three interconnected dimensions: context, production, and perception, namely by perceiving and responding to contextual cues to modulate both the selection of

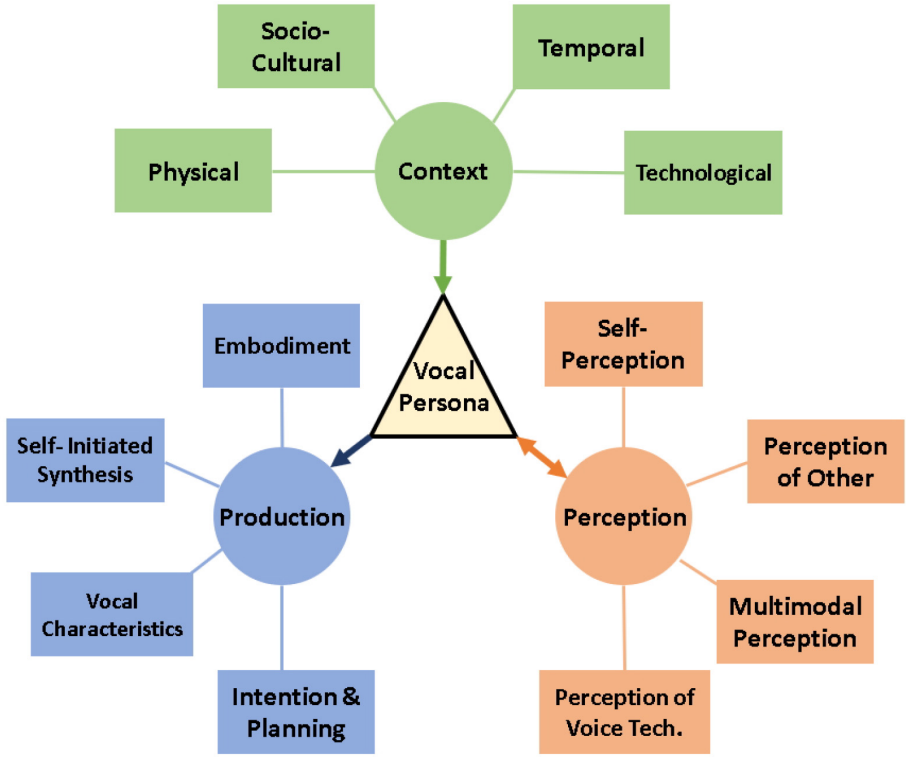

FIGURE 1
The proposed three-spoke model of vocal persona: Context, Production, and Perception. Each spoke represents a dimension of how vocal persona is constructed and understood. Arrows indicate bidirectional influence between dimensions. This model reflects the aggregation of insights from expert interviews and serves to pull together the many components of vocal persona. However, in implementation, its components could be incorporated at varying levels of granularity depending on downstream user needs.

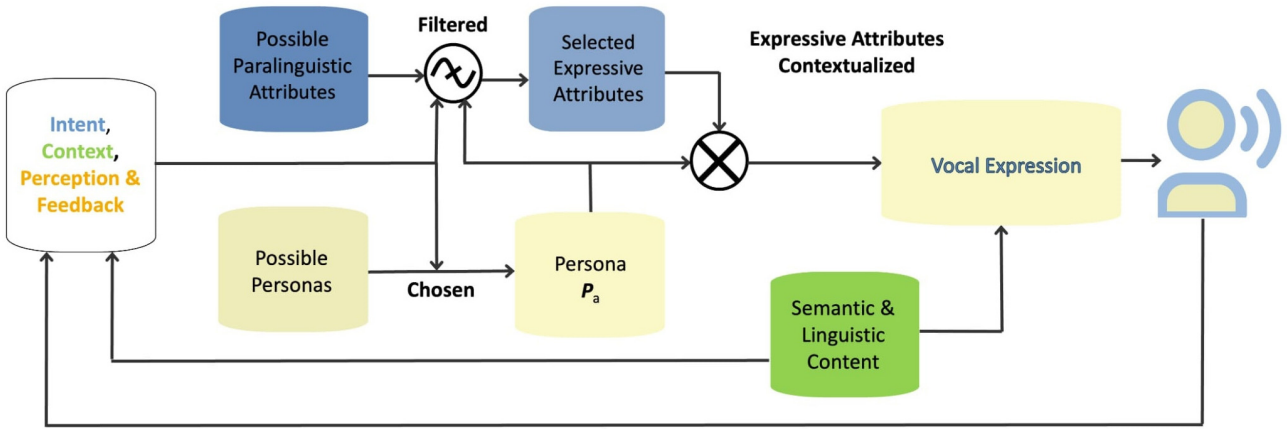

FIGURE 2
Flowchart illustrating persona-guided vocal expression.

vocal personas and the fine-grained paralinguistic adjustments within those personas. Participants emphasized that vocal choices are shaped not only by external environments but also by communicative goals, technical affordances, and perceived audience reception, supporting Goffman's notion of a shared "definition of the situation" (Goffman, 2016). These insights coalesced into the three-spoke model presented in Figure 1. This model corroborates the theories of Goffman (2016), Butler (1990), and Barnlund (1970). Together, these theoretical foundations, supported by data from our interviews, underscore the utility of the model in both interpreting and crafting adaptive, context-sensitive vocal personas for interactions between humans and between humans and computers. The model's three spokes work in tandem to guide the persona's development and ensure it is both relevant

and adaptable across communication scenarios. This leads us to a precise definition of a vocal persona: *"a chosen set of vocal expressions used to orient and respond to a communication context."*

Our findings show that there is a structured relationship between context, vocal expression, and identity that governs vocal persona in both natural interactions and VUI systems. Our analysis yielded a range of emergent themes within the umbrella of context, perception and production that not only corroborate established theories but also extend our understanding of vocal persona in today's rapidly evolving technological landscape. In line with the Interaction and Transaction Models of Communication (Westley and Malcolm S. MacLean, 1957; Barnlund, 1970), our results demonstrate that vocal behavior is inherently dynamic, reflecting a complex interplay between environmental, emotional, and sociocultural contexts. Participants consistently emphasized that vocal choices are part of a continuous sender–receiver dynamic, where both parties contribute to a shared perception, i.e., "definition of the situation" as described by Goffman (2016). This shared perception of context is critical; it underpins the selection and adaptation of vocal personas to ensure that expressions are effective and appropriate.

Our findings also suggest that vocal persona enactment often occurs as a holistic shift rather than as isolated adjustments to acoustic features. While participants, especially those with performance backgrounds, often reflected on specific vocal parameters in retrospect, they more commonly described transitioning between personas in embodied and intuitive ways. Even the performer subgroup, who possessed the ability to finely manipulate vocal elements, often defaulted to expressing entire personas fluidly, guided by context or emotional stance rather than deliberate control. This continuum—from unconscious, adaptive modulation to more deliberate adjustments—underscores that the process is not one of toggling between discrete, bounded personas, but rather of navigating a dynamic and flexible expressive space, often in service of maintaining authenticity, coherence, self-perception, or alignment with internal state while responding to context.

5.1 Design insights and future research directions

Finally, our findings yield actionable insights for improving expression and personalization in Voice User Interfaces (VUIs) and Augmentative and Alternative Communication (AAC) systems.

5.1.1 Prioritization of paralinguistic integrity

Participants described how the alignment between paralinguistic cues and their communicative intent shaped perceived authenticity and trust. Even under technically constrained conditions (e.g., phone calls), the perceived congruence between voice and intent mattered more than fidelity. This finding challenges prevailing design paradigms that prioritize audio quality over expressive integrity, and underscores the importance of designing systems that can preserve nuanced social signals.

5.1.2 Agency in AAC device interaction

Vocal persona is not just about sound but about conveying intentionality. Participants emphasized the importance of controlling whether and how emotional states are reflected in voice. AAC systems should support this agency and explicitly indicate the source of any misalignment, whether stemming from the user, intermediary, or system.

5.1.3 Flexible voice attribute description

Users described vocal attributes using both technical and affective language. This suggests a need for interfaces that allow engagement at multiple levels of abstraction. Descriptors should support natural variation across users while still mapping to controllable acoustic dimensions.

5.1.4 Multimodal awareness and feedback

Perception—as defined in our model—can be inferred from indirect signals such as user correction behavior, longitudinal interaction patterns, or feedback. Incorporating multimodal sensors (e.g., ambient noise detection, haptics, or facial expression tracking) could help VUIs respond dynamically and contextually. For example, real-time adaptation of vocal intensity in noisy environments or haptic indicators of voice modulation could enhance embodiment and agency.

5.1.5 Data-driven adaptation to context

Future systems should incorporate a wide range of contextual signals—physical, social, affective—to construct rich conditioning vectors for vocal synthesis. Advances in multimodal learning and self-supervised representation make this increasingly feasible. By treating context as a first-class input, systems can deliver nuanced and situationally appropriate vocal outputs.

5.1.6 Embedding persona controls in real-world design pipelines

In the current era of expressive AI models, our framework could be used to address a critical gap. Prompt-based voice synthesis (e.g., PromptTTS Guo et al., 2022) enables broad stylistic control, while example-based or parameter-based systems (Valle et al., 2020b; Wang et al., 2018; Hsu et al., 2019; Morrison et al., 2021a) allow fine-tuned manipulation. Yet these paradigms remain largely unintegrated. By centering the dimensions of context, production, and perception, our model provides a structure for connecting high-level persona prompts with more granular paralinguistic control—a union that mirrors natural human expressivity.

As such, VUI should offer layered, modular interactions that integrate high-level persona prompting with granular paralinguistic control, enabling both expert-level customization and simplified interfaces for casual users. We see this layering as a key implication for both interaction design and in the design of neural models powering vocal synthesis. The model's full expressivity can be exposed in expert-facing tools (e.g., for

co-design), while simplified versions (e.g. style synthesis through a priori prompting) can be offered to casual users. Controls such as sliders or persona prompts could be valuable in co-design sessions or annotation schemes, particularly with AAC practitioners and users. These affordances might help clarify communicative intent, enable customization, or serve as scaffolding for new voice learners. Neural network architectures might allow high-level natural language prompts to be translated into conditioning tokens, with deeper paralinguistic controls made accessible as needed—an approach aligned with foundational HCI work on user-tailorable systems (Carroll and Rosson, 1987; Fischer, 2001) and with controllable TTS research (Valle et al., 2020a). Future work should evaluate how such tools might integrate into existing designs and workflows.

5.2 Limitations and future work

Several limitations must be acknowledged. Our participant sample skewed toward Western and professional contexts, which may limit the generalizability of the model. Non-aural modalities such as signed languages were not addressed, and the model may require adaptation in multimodal or multilingual settings.

Although our framework provides a structured view of vocal persona, applying it in computational systems presents open challenges (Valencia et al., 2020; Pullin et al., 2017). Bridging intuitive human control with programmable interface design requires thoughtful abstraction. Future systems may draw inspiration from domains like digital musical instruments (DMIs) or audio workstations (DAWs), offering layered control to accommodate both novice and expert users.

A major direction for future work is empirical evaluation. While this study focused on expert users in order to capture the full depth of vocal persona construction, assessing which aspects of the model transfer effectively to broader populations and user groups will be critical. Future empirical studies should examine how much control should be exposed, what abstractions feel intuitive, and how interfaces can support flexible, adaptive, and accessible vocal expression across user groups and use cases.

6 Conclusion

This study re-frames vocal expression through the lens of *vocal persona*, a context-sensitive set of vocal behaviors that integrates broad identity cues with fine-grained paralinguistic modulation. Through thematic analysis of interviews with vocal professionals, we found that persona is enacted through dynamic shifts across multiple expressive layers, guided by communicative intent and contextual demands.

We propose a three-spoke model of vocal persona—*Context*, *Production*, and *Perception*—that captures the interplay between situational factors, technical control, and social interpretation. This model supports a persona-guided framework that bridges high-level intent and planning with paralinguistic nuance. In doing so, it helps unify previously separate approaches in expressive speech synthesis: prompt-based persona shaping and low-level acoustic control.

Finally, our findings offer actionable insights for the design of Voice User Interfaces (VUIs) and Augmentative and Alternative Communication (AAC) systems. Emphasizing user agency, multimodal awareness, and data-driven contextual adaptation, this work lays the groundwork for more expressive, responsive, and user-centered voice technologies.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

The studies involving humans were approved by the Stanford University Research Compliance Office. The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study. Written informed consent was obtained from the individual(s) for the publication of any potentially identifiable images or data included in this article.

Author contributions

CN: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Resources, Visualization, Writing – original draft, Writing – review & editing. LM: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Resources, Visualization, Writing – review & editing. JB: Funding acquisition, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. Funding for participants was provided by a Ric Weiland Graduate Fellowship Grant.

Acknowledgments

The authors thank Drs. So Yeon Park and Noah Fram for their guidance and feedback on study design.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declare that no Gen AI was used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcomp.2025.1575296/full#supplementary-material

References

Addington, D. W. (1968). The relationship of selected vocal characteristics to personality perception. *Speech Monogr.* 35, 492–503. doi: 10.1080/03637756809375599

Bachorowski, J. A. (1999). Vocal expression and perception of emotion. *Curr. Dir. Psychol. Sci.* 8, 53–57. doi: 10.1111/1467-8721.00013

Barnlund, D. C. (1970). "A transactional model of communication," in *Language Behavior* (De Gruyter Mouton), 47–57. doi: 10.4324/9781315080918-5

Berlo, D. (1960). *The Process of Communication: An Introduction to Theory and Practice*. New York: Holt, Reinhart and Winston.

Bullock, B. E., and Toribio, A. J. (2009). *The Cambridge Handbook of Linguistic Code-Switching*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511576331

Butler, J. (1990). Gender trouble, feminist theory, and psychoanalytic discourse. *Feminism/Postmoder.* 327, 324–340.

Carroll, J. M., and Rosson, M. B. (1987). Paradox of the Active User. Cambridge, MA, USA: MIT Press, 80–111.

Fischer, G. (2001). User modeling in human-computer interaction. User Model. User-Adapted Inter. 11, 65–86. doi: 10.1023/A:1011145532042

Goffman, E. (2016). "The presentation of self in everyday life," in *Social Theory Re-Wired: New Connections to Classical and Contemporary Perspectives: Second Edition* (University of Edinburgh), 482–493.

Guldner, S., Lavan, N., Lally, C., Wittmann, L., Nees, F., Flor, H., et al. (2024). Human talkers change their voices to elicit specific trait percepts. *Psychon. Bull. Rev.* 31, 209–222. doi: 10.3758/s13423-023-02333-y

Guo, Z., Leng, Y., Wu, Y., Zhao, S., and Tan, X. (2022). "PromptTTS: controllable text-to-speech with text descriptions," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE), 1–5. doi: 10.1109/ICASSP49357.2023.10096285

Hsu, W. N., Zhang, Y., Weiss, R. J., Zen, H., Wu, Y., Wang, Y., et al. (2019). "Hierarchical generative modeling for controllable speech synthesis," in *7th International Conference on Learning Representations, ICLR 2019*.

Linklater, K. (2006). *Freeing the Natural Voice: Imagery and Art in the Practice of Voice and Language*. Hollywood, CA: Drama Publishers.

Marshall, P. D., and Barbour, K. (2015). Making intellectual room for persona studies: a new consciousness and a shifted perspective. *Persona Stud.* 1, 1–12. doi: 10.21153/ps2015vol1no1art464

McAleer, P., Todorov, A., and Belin, P. (2014). How do you say 'hello'? Personality impressions from brief novel voices. *PLoS ONE* 9:90779. doi: 10.1371/journal.pone.0090779

Morrison, M., Jin, Z., Bryan, N. J., Caceres, J.-P., and Pardo, B. (2021a). Neural pitch-shifting and time-stretching with controllable LPCNet. *arXiv preprint arXiv:2110.02360*.

Morrison, M., Rencker, L., Jin, Z., Bryan, N. J., Caceres, J. P., and Pardo, B. (2021b). "Context-aware prosody correction for text-based speech editing," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings 2021-June*, 7038–7042. doi: 10.1109/ICASSP39728.2021.9414633

Nass, C., Moon, Y., Reeves, B., and Dryer, C. (1995). "Can computer personalities be human personalities?" in *Conference Companion on Human Factors in Computing Systems*, 228–229. doi: 10.1145/223355.223538

Nass, C., Steuer, J., and Tauber, E. R. (1994). "Computers are social actors," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 72–78. doi: 10.1145/191666.191703

Nass, C. I., Moon, Y., and Reeves, B. (2001). Does computer-synthesized speech manifest personality? Experimental tests of recognition, similarity-attraction, and consistency-attraction. *J. Exper. Psychol. Appl.* 7, 27–42. doi: 10.1037/1076-898X.7.3.171

Neekhara, P., Hussain, S., Dubnov, S., Koushanfar, F., and McAuley, J. (2021). "Expressive neural voice cloning," in *Asian Conference on Machine Learning* (PMLR), 252–267.

Nowell, L. S., Norris, J. M., White, D. E., and Moules, N. J. (2017). Thematic analysis: Striving to meet the trustworthiness criteria. Int. J. Qualit. Methods 16:1609406917733847. doi: 10.1177/1609406917733847

OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., et al. (2024). Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Pullin, G., and Hennig, S. (2015). 17 ways to say yes: toward nuanced tone of voice in AAC and speech technology. *Augment. Altern. Commun.* 31, 170–180. doi: 10.3109/07434618.2015.1037930

Pullin, G., Treviranus, J., Patel, R., and Higginbotham, J. (2017). Designing interaction, voice, and inclusion in AAC research. *AAC* 33, 139–148. doi: 10.1080/07434618.2017.1342690

Scherer, K. R. (1978). Personality inference from voice quality: the loud voice of extroversion. *Eur. J. Soc. Psychol.* 8, 467–487. doi: 10.1002/ejsp.2420080405

Schuller, B., Steidl, S., Batliner, A., Burkhardt, F., Devillers, L., Müller, C., et al. (2013). Paralinguistics in speech and language—state-of-the-art and the challenge. *Comput. Speech Lang.* 27, 4–39. doi: 10.1016/j.csl.2012.02.005

Shechtman, S., Fernandez, R., Sorin, A., and Haws, D. (2021). "Synthesis of expressive speaking styles with limited training data in a multi-speaker, prosody-controllable sequence-to-sequence architecture," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 3456–3460. doi: 10.21437/Interspeech.2021-1446

Shimizu, R., Yamamoto, R., Kawamura, M., Shirahata, Y., Doi, H., Komatsu, T., et al. (2024). "Prompttts++: controlling speaker identity in prompt-based text-to-speech using natural language descriptions," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 12672–12676. doi: 10.1109/ICASSP48485.2024.10448173

Sorin, A., Shechtman, S., and Rendel, A. (2017). "Semi parametric concatenative TTS with instant voice modification capabilities," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 1373–1377. doi: 10.21437/Interspeech.2017-1202

Stern, J., Schild, C., Jones, B. C., DeBruine, L. M., Hahn, A., Puts, D. A., et al. (2021). Do voices carry valid information about a speaker's personality? *J. Res. Pers.* 92:104092. doi: 10.1016/j.jrp.2021.104092

Tagg, P. (2012). "Vocal persona," in *Music's Meanings*.

Terry, G., Hayfield, N., Clarke, V., and Braun, V. (2017). "Thematic analysis," in *The SAGE Handbook of Qualitative Research in Psychology*, 17–37. doi: 10.4135/9781526405555.n2

Valencia, S., Pavel, A., Santa Maria, J., Yu, S., Bigham, J. P., and Admoni, H. (2020). "Conversational agency in augmentative and alternative communication," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–12. doi: 10.1145/3313831.3376376

Valle, R., Li, J., Prenger, R., and Catanzaro, B. (2020a). "Mellotron: multispeaker expressive voice synthesis by conditioning on rhythm, pitch and global style tokens," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings* (Institute of Electrical and Electronics Engineers Inc.), 6189–6193. doi: 10.1109/ICASSP40776.2020.9054556

Valle, R., Shih, K., Prenger, R., and Catanzaro, B. (2020b). FLOWTRON: an autoregressive flow-based generative network for text-to-speech synthesis. *arXiv preprint arXiv:2005.05957*.

Wang, C., Chen, S., Wu, Y., Zhang, Z., Zhou, L., Liu, S., et al. (2023). Neural codec language models are zero-shot text to speech synthesizers. *arXiv preprint arXiv:2301.02111*.

Wang, Y., Stanton, D., Zhang, Y., Skerry-Ryan, R. J., Battenberg, E., Shor, J., et al. (2018). "Style tokens: unsupervised style modeling, control and transfer in end-to-end speech synthesis," in *35th International Conference on Machine Learning, ICML 2018*, 8229–8238.

Westley, B. H., and Malcolm, S., MacLean, J. (1957). A conceptual model for communications research. *Journalism Quart.* 34, 31–38. doi: 10.1177/107769905703400103

Yang, D., Liu, S., Huang, R., Weng, C., and Meng, H. (2024). Instructtts: modelling expressive tts in discrete latent space with natural language style prompt. *IEEE/ACM Trans. Audio, Speech, Lang. Proc.* 32, 2913–2925. doi: 10.1109/TASLP.2024.3402088